METHOD AND SYSTEM FOR GENERATING FACIAL ANIMATION VALUES BASED ON A COMBINATION OF VISUAL AND AUDIO INFORMATION

BACKGROUND OF THE INVENTION

The present invention relates to avatar animation, and more particularly, to facial feature tracking.

Virtual spaces filled with avatars are an attractive the way to allow for the experience of a shared environment. However, animation of a photo-realistic avatar often requires tedious efforts to generate realistic animation information.

Accordingly, there exists a significant need for improved techniques for generating animation information. The present invention satisfies this need.

SUMMARY OF THE INVENTION

The present invention is embodied in a method, and related apparatus, for generating facial animation values using a sequence of facial image frames and synchronously captured audio data of a speaking actor. In the method, a plurality of visual facial animation values are provided based on tracking of facial features in the sequence of facial image frames of the speaking actor, and a plurality of audio facial animation values are provided based on visemes detected using the synchronously captured audio voice data of the speaking actor. The plurality of visual facial animation values and the plurality of audio facial animation values are combined to generate output facial animation values for use in facial animation.

In more detailed features of the invention, the output facial animation values associated with a mouth for a facial animation may be based only on the respective mouth-associated values of the plurality of audio facial animation values. Alternatively, the output facial animation values associated with a mouth for a facial animation may be based on a weighted average of the respective mouth-associated values of the plurality of visual facial animation values and the respective mouth-associated values of the plurality of audio facial animation values. Also, the output facial animation values associated with a mouth for a facial animation may be based on Kalman filtering of the respective mouth-associated values of the plurality of visual facial animation values and the respective mouth-associated values of the plurality of audio facial

animation values. Further, the step of combining the plurality of visual facial animation values and the plurality of audio facial animation values to generate output facial animation values may include detecting whether speech is occurring in the synchronously captured audio voice data of the speaking actor and, while speech is detected as occurring, generating the output facial animation values associated with a mouth based only on the respective mouth-associated values of the plurality of audio facial animation values and, while speech is not detected as occurring, generating the output facial animation values associated with the mouth based only on the respective mouth-associated values of the plurality of visual facial animation values.

In other more detailed features of the invention, the tracking of facial features in the sequence of facial image frames of the speaking actor may be performed using bunch graph matching, or using transformed facial image frames generated based on wavelet transformations, such as Gabor wavelet transformations, of the facial images.

Other features and advantages of the present invention should be apparent from the following description of the preferred embodiments taken in conjunction with the accompanying drawings, which illustrate, by way of example, the principles of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a flow diagram for illustrating a method for generating facial animation values using a sequence of facial image frames and synchronously captured audio data of a speaking actor, according to the present invention.

FIG. 2 is a flow diagram for illustrating a technique for combining visual facial animation values and audio facial animation values, according to the present invention.

FIG. 3 is a block diagram for illustrating a technique for selectively combining visual facial animation values and audio facial animation values, according to the present invention.

FIG. 4 is a block diagram of a technique for detecting speech activity in audio data.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention is embodied in a method, and related apparatus, for generating facial animation values using a sequence of facial image frames and synchronously captured audio data of a speaking actor.

As shown in FIG. 1, the method includes providing a plurality of visual-facial-animation values v_n (step 12) based on tracking of facial features in the sequence of facial image frames of the speaking actor (step 14), and providing a plurality of audio-facial-animation values a_n (step 16) based on visemes detected using the synchronously captured audio voice data of the speaking actor (step 18). The plurality of visual facial animation values and the plurality of audio facial animation values are combined to generate output facial animation values f_n for use in facial animation (step 20).

The output facial animation values associated with a mouth in the facial animation may be based only on the respective mouth-associated values of the plurality of audio facial animation values. The combination of the visually generated facial animation values and the audio-based mouth animation values provides advantageous display of animated avatars.

The visemes are a visual equivalent of phonemes, i.e., visemes are related to facial expressions that are associated with temporal speech units in audio voice data. For the English language, it is generally agreed that there may be 15 visemes associated with 43 possible phomenes. Speech analysis and viseme detection may be accomplished with analysis products produced by LIPSinc, Inc., of Morrisville, North Carolina (www.lipsinc.com).

The facial animation values or tags may be displacement values relative to neutral face values. Advantageously, 8 to 22 (or more) facial animation values may be used to define and animate the mouth, eyes, eyebrows, nose, and the head angle. Representative facial animation values for the mouth may include vertical mouth position, horizontal mouth position, mouth width, lip distance, and mouth corner position (left and right).

With reference to FIG. 2, the output facial animation values f_n associated with a mouth in the facial animation may be based on a weighted average (Equation 1) of the respective mouth-associated values of the plurality of visual facial animation values v_n and the respective mouth-associated values of the plurality of audio facial animation values a_n (step 22'). The visual facial animation values and the audio facial animation values may be assigned weights, σ^v and σ^a, respectively, that are proportional to an uncertainty of the animation values. The merging of the values may be memoryless, i.e., each combined value may be the result of present values, exclusively.

$$\left(\underline{f}_n = \frac{\underline{\sigma}_n^a}{\underline{\sigma}_n^a + \underline{\sigma}_n^v} \cdot \underline{a}_n + \frac{\underline{\sigma}_n^v}{\underline{\sigma}_n^v + \underline{\sigma}_n^a} \cdot \underline{v}_n \right)_i \qquad \text{Equation. 1}$$

Alternatively, the combined values may be based on recursive estimates using a series of the animation values. Accordingly, the output facial animation values associated with a mouth in the facial animation are based on Kalman filtering of the respective mouth-associated values of the plurality of visual facial animation values and the respective mouth-associated values of the plurality of audio facial animation values. The Kalman filtering may be accomplished in accordance with Equations 2-7.

$$\underline{f}_n = \underline{\underline{A}}_{n-q} \cdot \underline{f}_{n-q} \qquad \text{Equation. 2}$$

$$\hat{\underline{f}}_n = \underline{f}_n + \underline{\underline{K}}_n \left[\frac{(\underline{a}_n - \underline{f}_n)}{(\underline{v}_n - \underline{f}_n)} \right] \qquad \text{Equation. 3}$$

$$\underline{\underline{\tilde{p}}}_n^f = \underline{\underline{A}}_{n-q} \cdot \underline{\underline{\tilde{p}}}_{n-a}^f \cdot \underline{\underline{A}}_{n-q} + \underline{\underline{Q}}_{n-q} \qquad \text{Equation. 4}$$

$$\underline{\underline{K}}_n = \underline{\underline{\tilde{p}}}_n^f \left(\underline{\underline{\tilde{p}}}_n^f + \underline{\underline{R}}_n \right)^{-q} \qquad \text{Equation. 5}$$

$$\underline{m}_n = \underline{f}_n \qquad \text{Equation. 6}$$

$$\hat{\underline{\underline{p}}}_n = \left(\underline{\underline{I}} - \underline{\underline{K}} \right) \cdot \underline{\underline{\tilde{p}}}_n^f \qquad \text{Equation. 7}$$

With reference to FIG. 3, the step of combining the plurality of visual facial animation values and the plurality of audio facial animation values to generate output facial animation values may include detecting whether speech is occurring in the synchronously captured audio voice data of the speaking actor. While speech is detected as occurring, the output facial animation values associated with a mouth may be generated based only on the respective mouth-

associated values of the plurality of audio facial animation values (switch S1 open, switch S2 closed). While speech is not detected as occurring, the output facial animation values associated with a mouth may be generated based only on the respective mouth-associated values of the plurality of visual facial animation values (S1 closed, S2 open).

The switches, S1 and S2, may be controlled by a Speech Activity Detector 22 (SAD). The operation of the SAD is described with reference to FIG. 4. The audio voice data 24 is filtered by a low-pass filter (step 26), and the audio features are computed for separating speech activity from background noise (step 28). The background noise may be characterized to minimize its effect on the SAD. The noise and audio speech indications 30 are temporally smoothed to decrease the effects of spurious detections of audio speech. (step 32)

The tracking of facial features in the sequence of facial image frames of the speaking actor may be performed using bunch graph matching, or using transformed facial image frames generated based on wavelet transformations, such as Gabor wavelet transformations, of the facial image frames. Wavelet-based tracking techniques are described in U.S. patent number 6,272,231. The entire disclosure of U.S. patent number 6,272,231 is hereby incorporated herein by reference. The techniques of the invention may be accomplished using generally available image processing systems.

Although the foregoing discloses the preferred embodiments of the present invention, it is understood that those skilled in the art may make various changes to the preferred embodiments without departing from the scope of the invention. The invention is defined only by the following claims.